Machine Learning Model to predict Maximum Heart Rate

Fahim Hoq

Content

Abstract

Cleveland dataset of 303 people was used to predict maximum heart rate from other predictors such as resting blood pressure, age, gender, serum cholesterol etc. Three different machine learning models such as multiple linear regression, random forest, BART was used to answer this question. Other variables were also taken into both models as confounding and precision variables. In all cases, root mean square error and R-squared was used to find the best model. In all 3 methods, ratio of training to test data set was 80 to 20.

Introduction

"The Rhythm of the heart has not only fascinated cardiologists but also inspired poets and musicians." (Stauss H.M. 2000.) Heart rate variability is also related with cardiovascular disease, in this paper, we are modeling how the heart rate is effected by blood pressure and how its effected by cholesterol. Since, blood pressure is an important predictors for cardiovascular mortality (Kikuya et. al 2000). Similarly, serum cholesterol having an effect on heart disease (Kannel, W. B. 1995). This paper tries to find the best model for predicting heart rate with machine learning techniques using the dataset of 303 Cleveland heart patients. This questions is important since understanding the predictors for maximum heart rate might allow people to take more preventative measures in order to decrease their chance of developing heart failure by trying to better control their blood pressure, and cholesterol levels for instance.

Data Description

The dataset was downloaded from kaggle. The data was collected by Robert Detrano, M.D., Ph.D of the Cleveland Clinic Foundation. The dataset has 14 variables. The response variable analyzed in this paper is the Maximum heart rate achieved. Sample size in the dataset is 303. There are fourteen variables. There were only six missing values, two for the variable nuclear stress test and four for the variable number of major vessels. (R. Detrano 2020). 80% of the data was used to train the model, while the remaining 20% was used to test it. More details about the dataset can be found in the appendix section.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
1	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0

Software Used for Analysis

In this paper, the data was analyzed using R-studio version 4.2.2. Library such as ggplot2 was used in plotting most of the figures for explanatory analysis. The library dplyr was used in cleaning and organizing the data. Moreover, the library car was used to find the variance inflation factor. Finally, libraries such as knitr, table1, htmlTable was used in making the table used in this paper. For more information please refer to the appendix and reference in this paper.

Explanatory Analysis

Histograms showed four out of the five continuous variables were approximately normal. Only, “depression induced by exercise relative to rest” variable was very right skewed. Bar plot of sex indicated that there are more than twice the number of male participants in the study compared to female. Scatter plot of hear rate with age showed a slightly downward slope. Which is slightly counter intuitive, perhaps there are confounding effect there. As we would naturally, expect older patients to have higher risk of cardiovascular disease, and hence have more maximum hear rate. Box plots showed all the categorical variables are distributed evenly with respect to maximum heart rate. Heat maps didn’t reveal any alarmingly high correlation between the variables. The variables “exercise relative to rest”, “exercise induced angina”, and “peak exercise segment” showed bit more correlation as expected.

Method1: Multiple Linear Regression

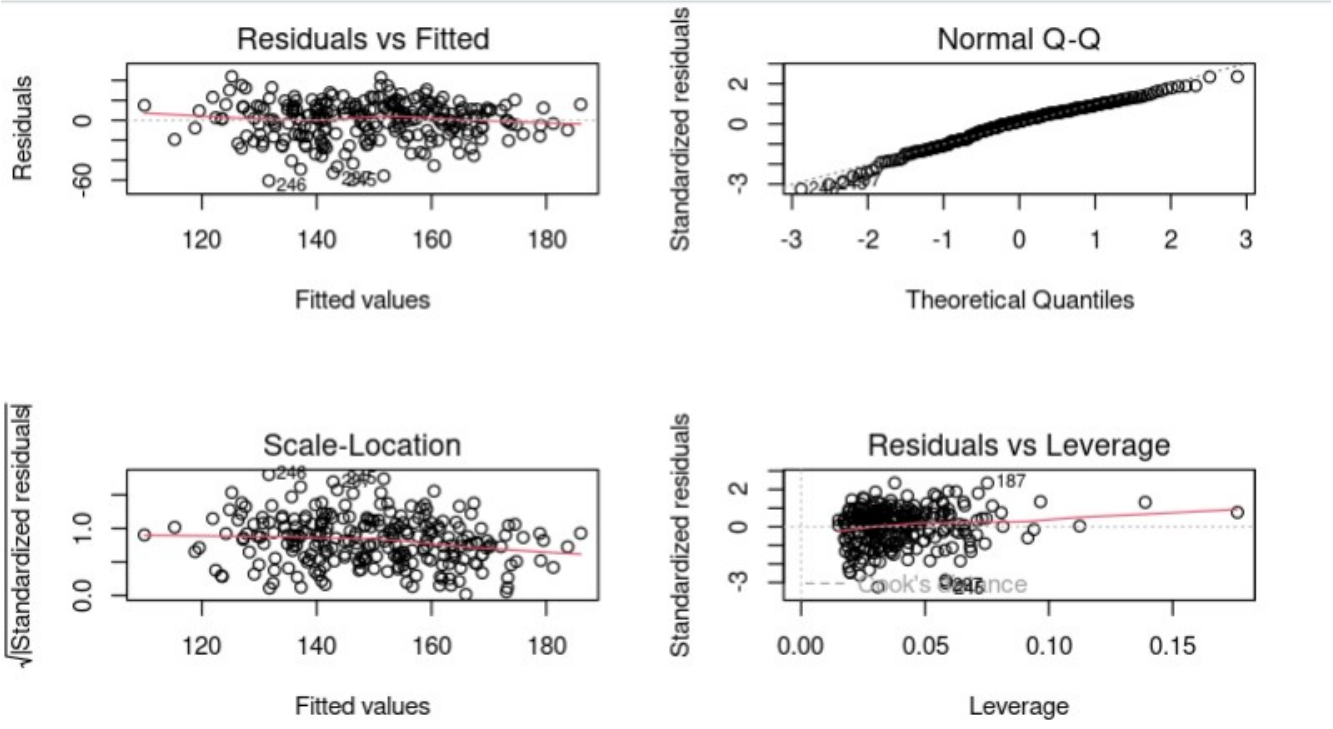
$y_i \sim B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4 + B_5x_5 + B_6x_6 + B_7x_7 + B_8x_8 + B_9x_9 + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$

Maximum Heart Rate (yi) ~ Intercept + chest pain(x_1) + age(x_2) + sex(x_3) + Blood sugar (x_4) + Blood Pressure(x_5) + Exercise Relative Rest(x_6) + Exercise Induced Angina(x_7) + Peak Exercise (x_8) + Cholesterol(x_9) + Error Term.

Advantage of linear model is that it is fairly robust, and it is easy to train and interpret. On the other hand, the disadvantage is that it is vulnerable to noise or over fitting, and is sensitive to outliers.

Assumptions and Diagnostic Plots

Assumptions of Linear models are the response or error term are independent and identically normally distributed with mean zero and constant variance. The diagnostic plot for both model shows the assumptions are satisfied. The residual verses fitted plot on both models indicates homoscedasticity as the points are spread almost evenly across. The quantile - quantile plot for both models are similar , they both show that the line is almost linear. The left side of the graph deviates slightly from straight but there is not a lot of data there. Hence, the normality assumption is satisfied. In the leverage plot, there does not appear to be any big outliers.

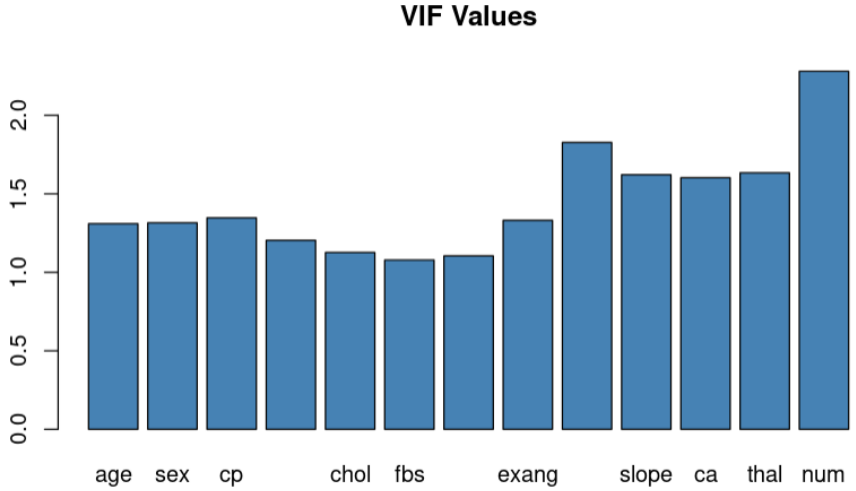


Multicollinearity

"Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another" (James G et. al 2021). Variance Inflation Factor (VIF_i) was used to find multicollinearity.

Where, $VIF_i = 1 / (1 - R_j^2)$, where R_j^2 is the adjusted R-squared.

The variance inflation factor for all the variables were below 5 for all the variables. Hence, multicollinearity is not a problem for this dataset either.

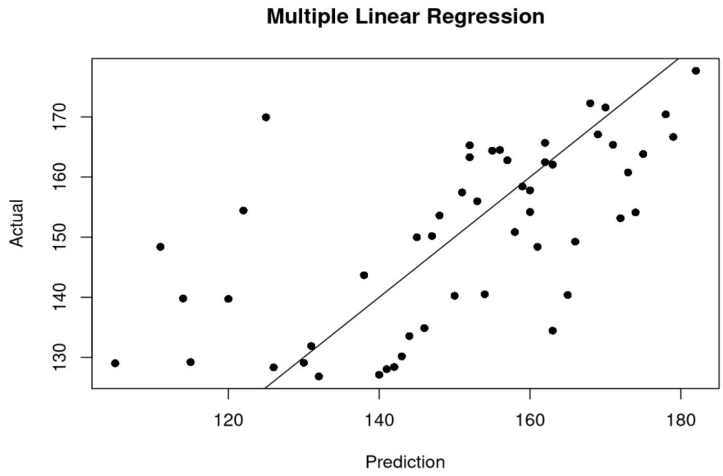


Results:

Summary of Multiple Linear Regression.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	205.973	12.727	16.184	0.000
age	-0.909	0.142	-6.418	0.000
sex	-0.799	2.694	-0.297	0.767
cp	-4.679	1.391	-3.363	0.001
trestbps	0.137	0.072	1.910	0.057
chol	0.035	0.024	1.445	0.150
fbs	0.014	3.440	0.004	0.997
exang	-8.859	2.923	-3.031	0.003
oldpeak	-1.871	1.330	-1.407	0.161
slope	-8.358	2.358	-3.545	0.000

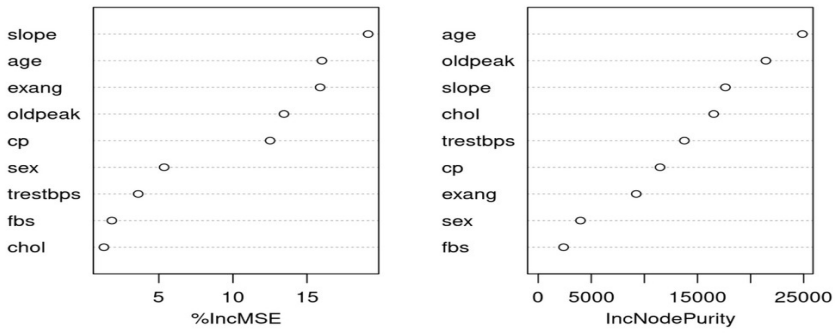
Here, we see that age, chest pain, exercise induced angina and peak exercise were significant factors for heart rate. We got an R-squared value of 0.3769 and root mean squared value of 14.84 for the multiple linear regression.



Method2: Random Forest

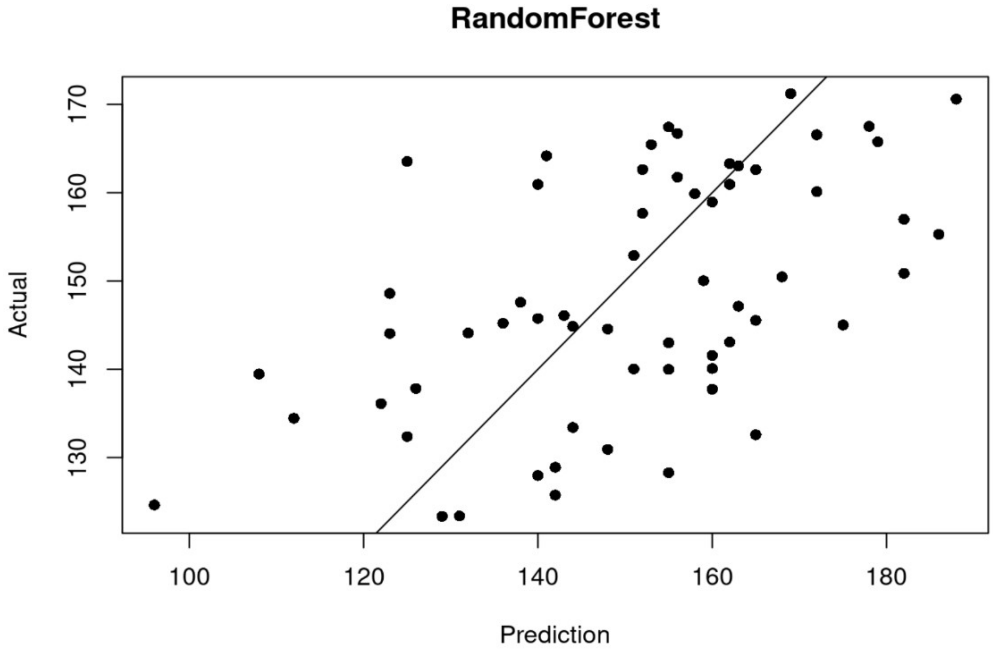
Bagging creates a huge number of decision trees, but in random forest takes only a subset of predictors. Usually, square root of predictors. In fact, that's what I have chosen for my analysis. Unlike Bagging, random forest have less variance due to forcing split to consider only a subset of predictors. Random Forest provides improvement over bagging by decorrelating the trees (James G et. al 2021).

So advantages of random forest is high accuracy, robustness, is not sensitive to missing values or outliers etc. But on the other hand, it is not easy to interpret, it uses lot of computer resources especially for big data. Thus takes long time to predict. It is also vulnerable to over fitting, if it captures noise in the data. The plot below shows the important variable for our model.



Results:

We got an R-squared value of 0.247 and root mean squared value of 16.19 for the multiple linear regression.



Method3: BART

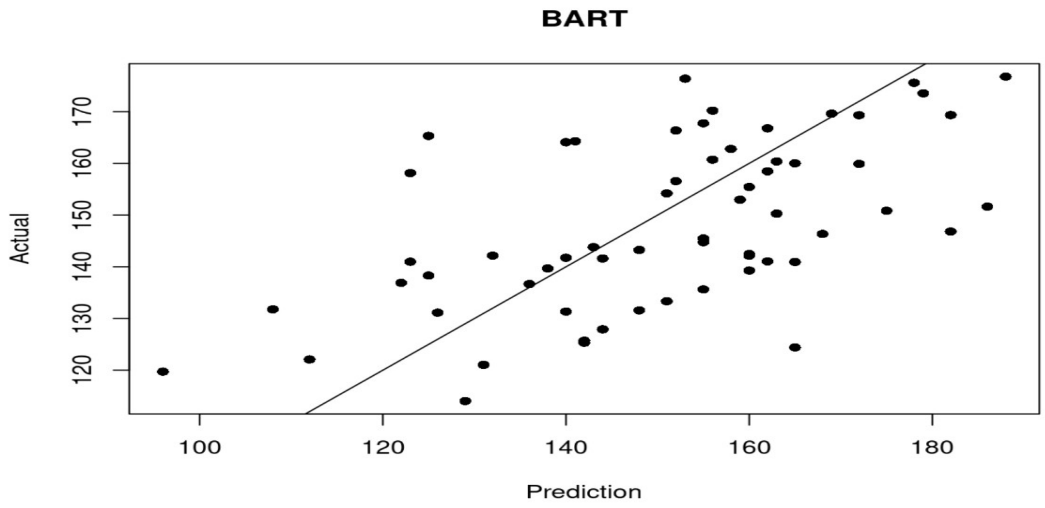
BART stands for Bayesian additive regression trees. As the name suggest it uses Bayesian method to estimate our parameters. So, it uses prior distribution to model the response with its

explanatory variable. Just like random forest, it also belongs in regression trees. Advantages of BART is that it can work with wide range of data. The method also happens to be robust to outliers and missing values. But on the other hand, it is usually hard to identify or characterize individual predictors that have strong influences on the response variable. (James G et. al 2021).

```
##       cp      sex      fbs    exang    slope trestbps      age  oldpeak
##   29.466   29.310   28.069   28.058   27.554   23.983   23.645   23.381
##     chol
##   15.229
```

Results

We got an R-squared value of 0.239 and root mean squared value of 17.97 for the multiple linear regression.



In all three methods, eighty percent of the data was used to train the model, and the other twenty percent was used to make predictions and test its accuracy.

Limitations of this study and Future Study

There are more than twice the number of male patients in the data compared to female. So even though gender is an important variable in random forest plot, it is hard to come to any conclusive statement regarding this variable. For future study, we might want to run the analysis again after collecting a dataset where the ratio of male and female participants is similar. Another limitation was that stylostic and diastolic blood pressure was not here in the data set. From literature review,we know this is an important variable, perhaps reading for those would have given our model even better fit. Furthermore, sample size of 303 might not be low, but ideally, we might want bigger dataset to train our model. Finally, other machine learning methods, which are not applied in this paper might also lead to interesting results in future study.

Furthermore, the dataset only has maximum heart rate achieved for patients, we need to also look at minimum heart rate achieved as well, as low heart rate per minute is also related to cardiovascular disease (Stauss H.M. 2000). Finally, this dataset did not have variables such as diabetes. So we might also want to include those variables in our future modeling for future study as these could have make an enormous difference.

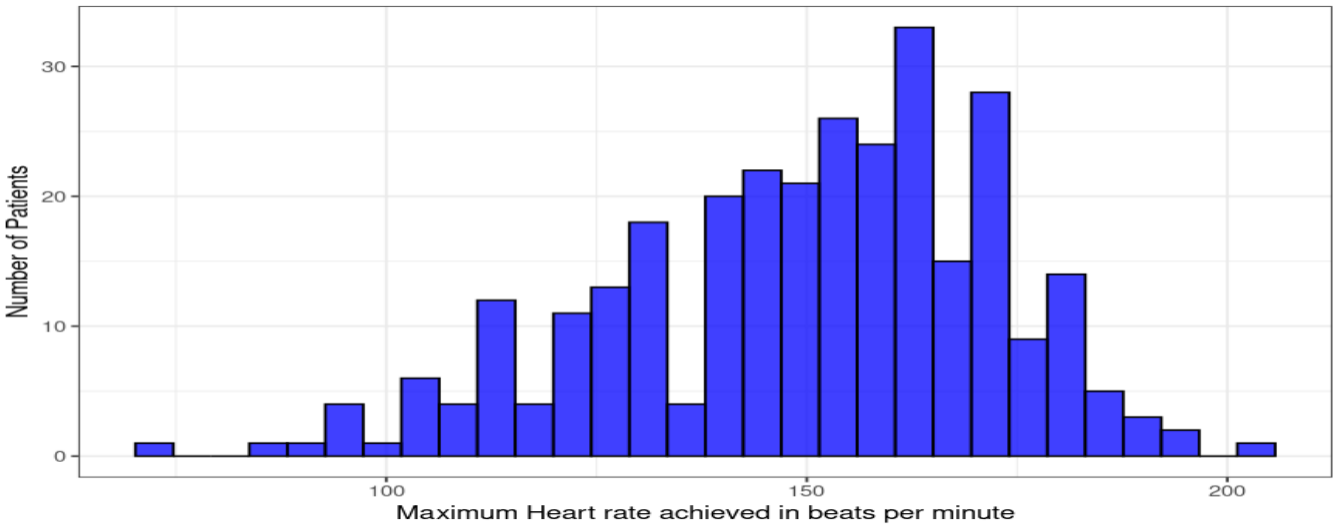
Conclusions & Interpretation

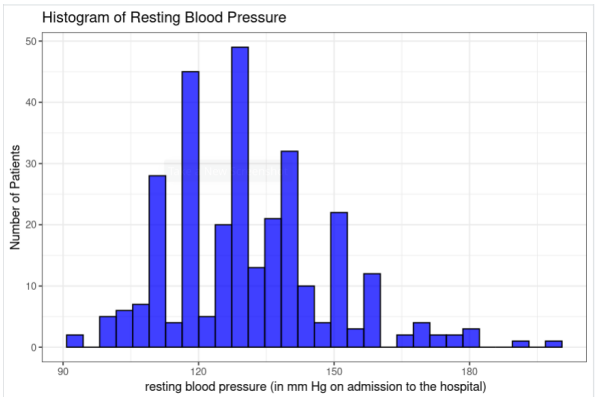
Methods	R-squared	Root Mean squared Error
Multiple Linear Regression	0.38	14.84
Random Forest	0.25	16.19
BART	0.24	17.97

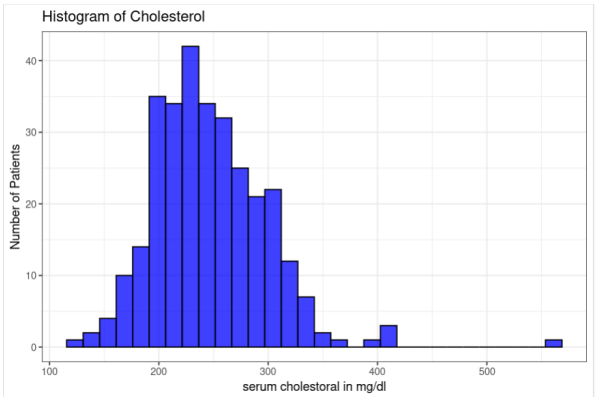
So, all three methods have low R-squared suggesting that none of them are necessarily an optimal fit for the dataset. In future study we might want to use other machine learning techniques to find even better model for our study. Having said that, from our three methods, Multiple Linear Regression performed the best, with the highest R-squared and the lowest root mean squared error.

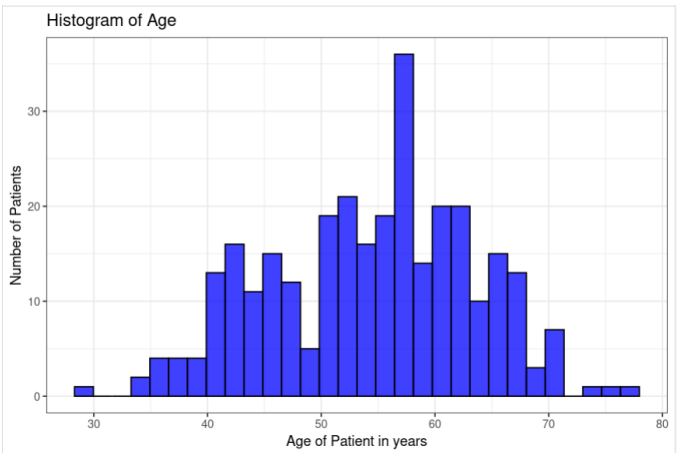
Appendix A (Plots)

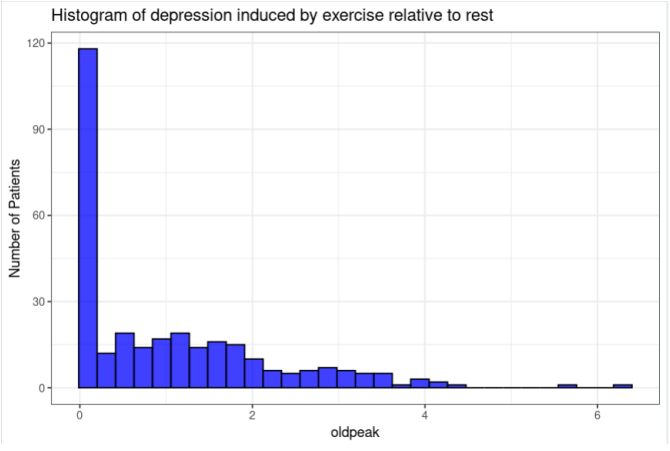
Histograms



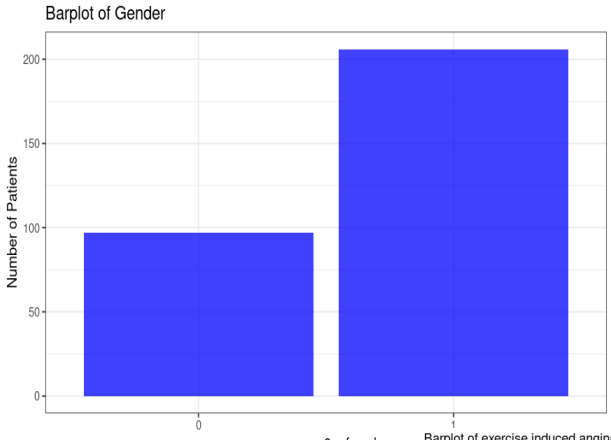


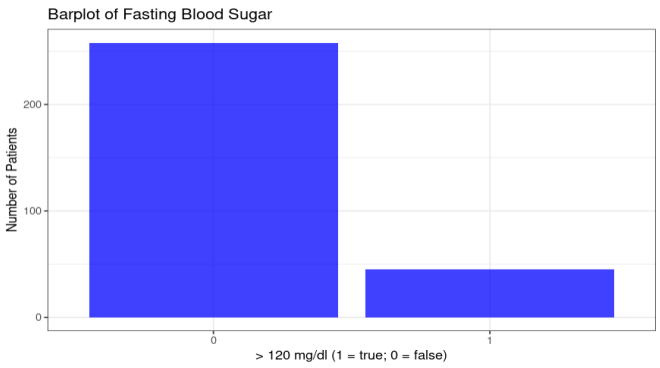


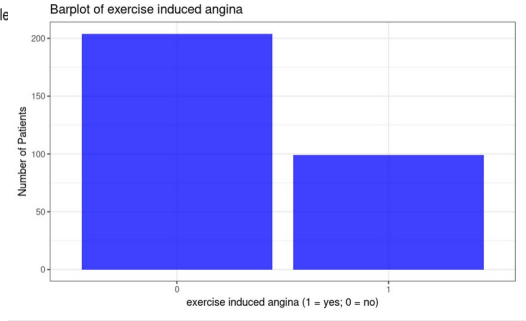


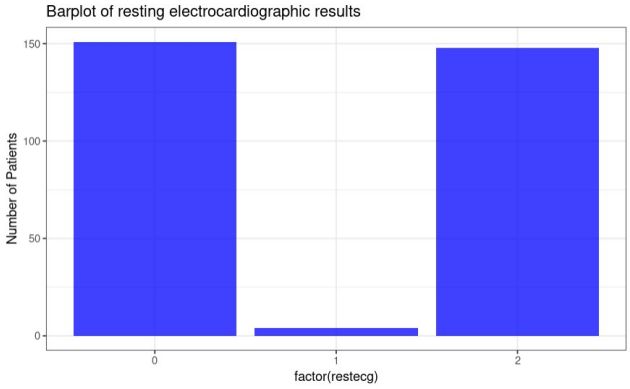


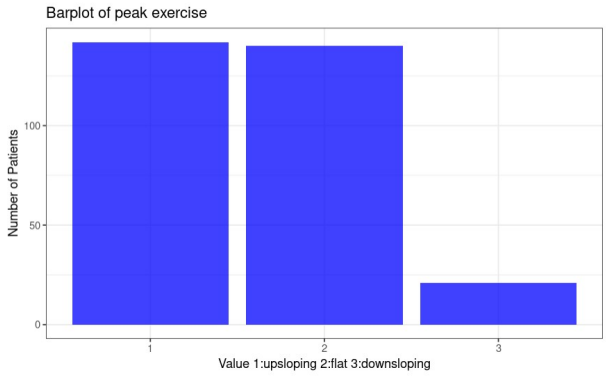
Bar Plots

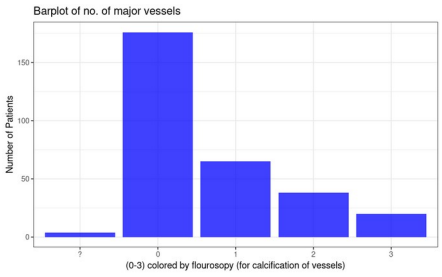

Barplot of Gender
Number of Patients
200
150
100
50
0
0
1

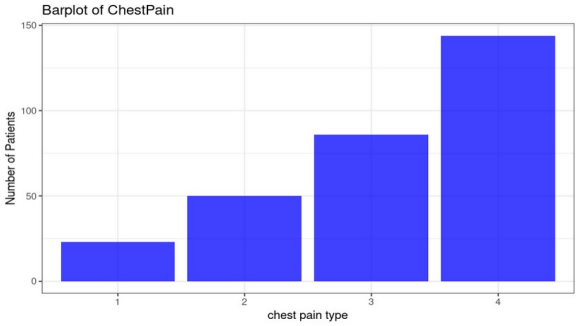

Barplot of Fasting Blood Sugar
Number of Patients
200
100
0
0
1
> 120 mg/dl (1 = true; 0 = false)


Barplot of exercise induced angina
Number of Patients
200
150
100
50
0
0
1
exercise induced angina (1 = yes; 0 = no)

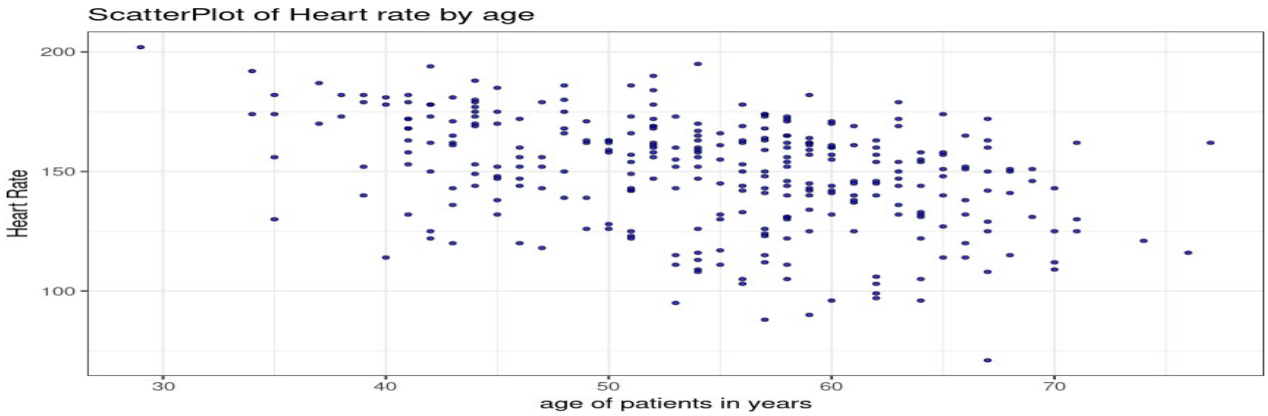

Barplot of resting electrocardiographic results
Number of Patients
150
100
50
0
0
1
2
factor(restecg)

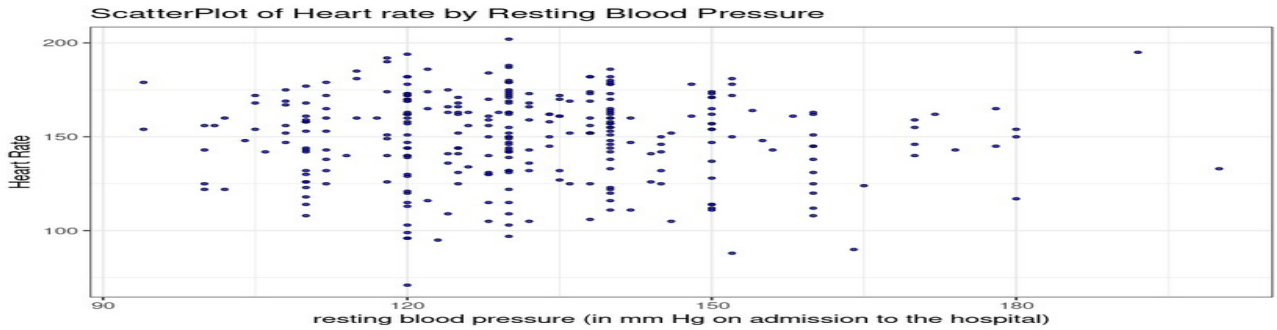

Barplot of peak exercise
Number of Patients
100
50
0
1
2
3
Value 1:upsloping 2:flat 3:downsloping

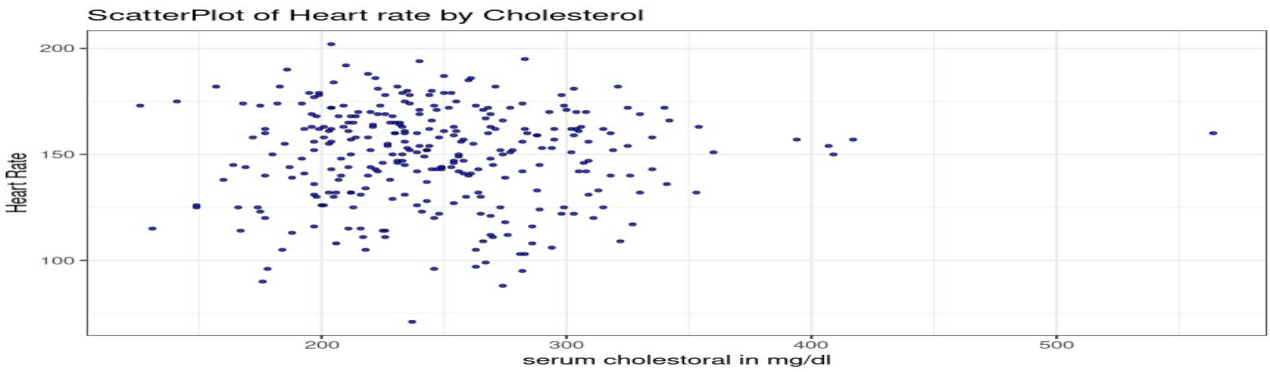

Barplot of no. of major vessels
Number of Patients
150
100
50
0
0
1
2
3
(0-3) colored by flourosopy (for calcification of vessels)

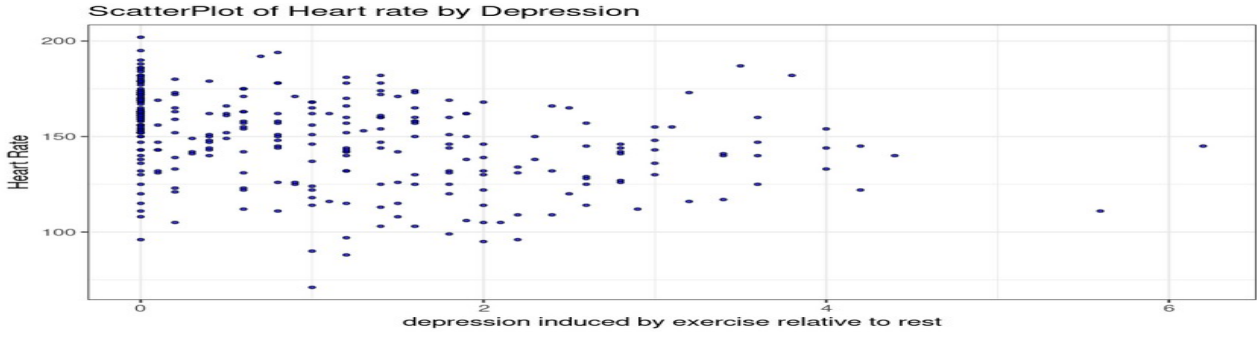

Barplot of ChestPain
Number of Patients
150
100
50
0
1
2
3
4
chest pain type

Scatter Plots

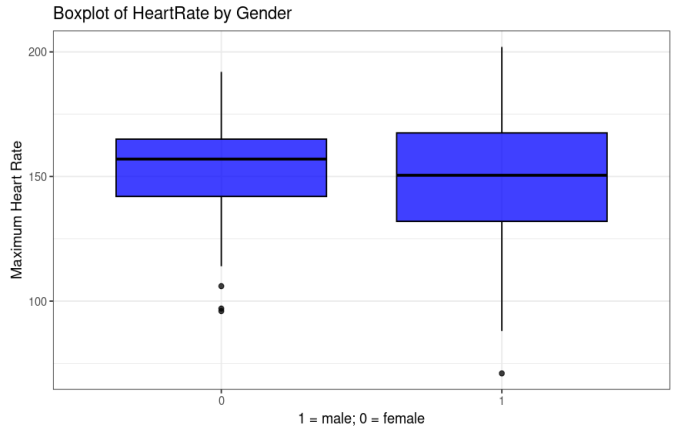


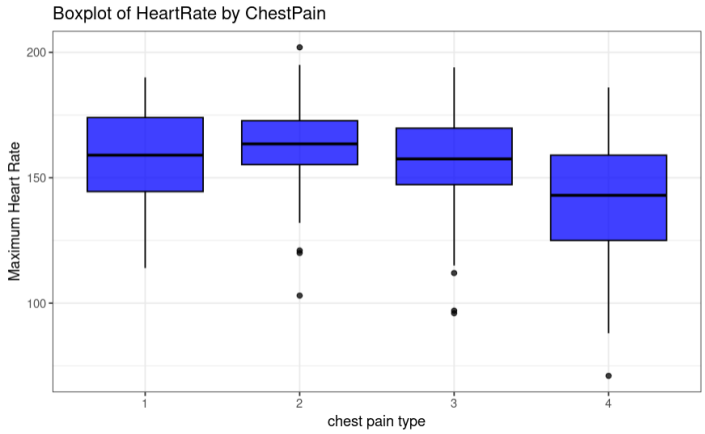


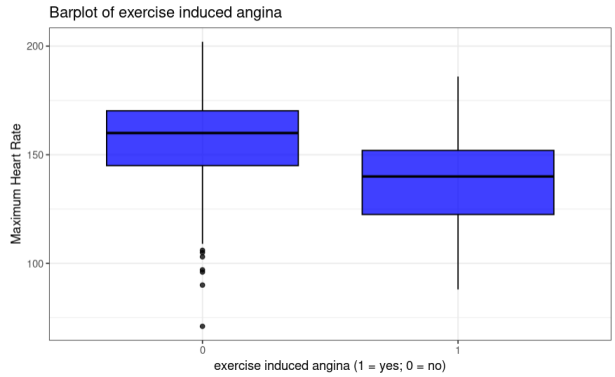


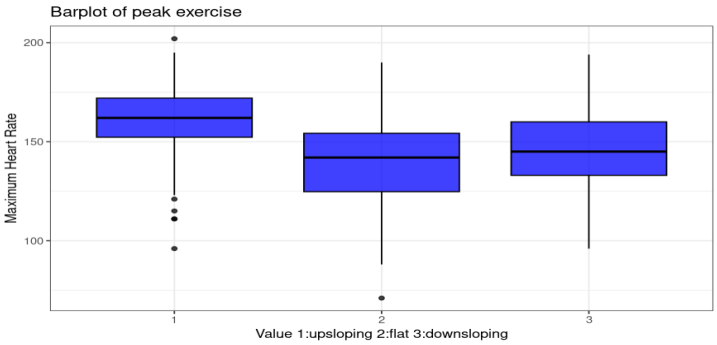


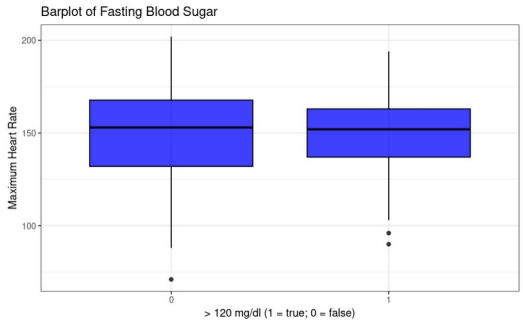
Box Plots

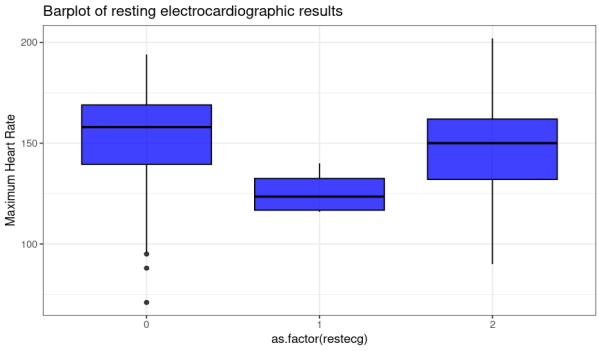

Boxplot of HeartRate by Gender
Maximum Heart Rate
200
150
100
0
1
1 = male; 0 = female

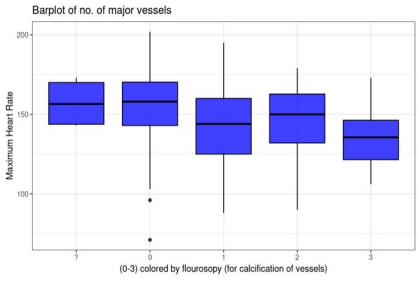

Boxplot of HeartRate by ChestPain
Maximum Heart Rate
200
150
100
1
2
3
4
chest pain type

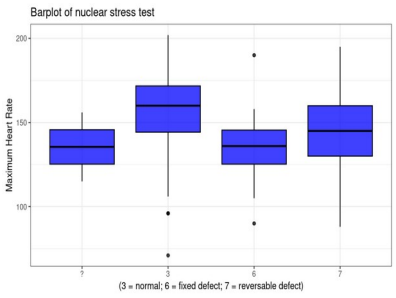

Barplot of exercise induced angina
Maximum Heart Rate
200
150
100
0
1
exercise induced angina (1 = yes; 0 = no)

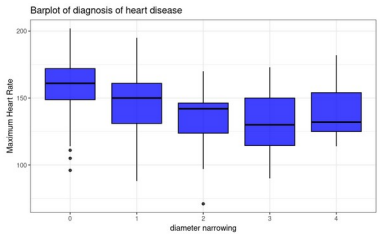

Barplot of peak exercise
Maximum Heart Rate
200
150
100
1
2
3
Value 1:upsloping 2:flat 3:downsloping


Barplot of Fasting Blood Sugar
Maximum Heart Rate
200
150
100
0
1
> 120 mg/dl (1 = true; 0 = false)


Barplot of resting electrocardiographic results
Maximum Heart Rate
200
150
100
0
1
2
as.factor(restecg)


Barplot of no. of major vessels
Maximum Heart Rate
200
150
100
0
1
2
3
(0-3) colored by flourosopy (for calcification of vessels)


Barplot of nuclear stress test
Maximum Heart Rate
200
150
100
3
6
7
(3 = normal; 6 = fixed defect; 7 = reversable defect)


Barplot of diagnosis of heart disease
Maximum Heart Rate
200
150
100
0
1
2
3
4
diameter narrowing

Appendix B (Tables)

Description of covariates

No.	Variable	Description
1	age	Age of Patents in years
2	sex	Gender of Patients
3	cp	Chest Pain Type (1: typical angina , 2: atypical angina, 3: non-anginal pain 4: asymptomatic)
4	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholestoral in mg/dl
6	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	resting electrocardiographic results (0: normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
12	ca	no. of major vessels (0-3) colored by flourosopy (for calcification of vessels)
13	thal	results of nuclear stress test (3: normal; 6: fixed defect; 7: reversable defect)
14	num	target variable representing diagnosis of heart disease (angiographic disease status) in any major vessel (0: < 50% diameter narrowing, 1: > 50% diameter narrowing)

Descriptive Statistics (Continuous Variables)

Summary of Descriptive Statistics

	Total (N=303)
Maximum Heart Rate achieved	
Mean (SD)	150 (22.9)
Median [Min, Max]	153 [71.0, 202]
Age of Patients	
Mean (SD)	54.4 (9.04)
Median [Min, Max]	56.0 [29.0, 77.0]
Resting Blood Pressure in mm Hg	
Mean (SD)	132 (17.6)
Median [Min, Max]	130 [94.0, 200]
serum Cholesterol in mg/dl	
Mean (SD)	247 (51.8)
Median [Min, Max]	241 [126, 564]
depression induced by exercise	
Mean (SD)	1.04 (1.16)
Median [Min, Max]	0.800 [0, 6.20]

Descriptive Statistics (Categorical Variables)

Summary of Descriptive Statistics

	Total (N=303)
Gender	
Female	97 (32.0%)
Male	206 (68.0%)
Chest Pain	
typical angina	23 (7.6%)
atypical angina	50 (16.5%)
non-anginal pain	86 (28.4%)
asymptomatic	144 (47.5%)
Fasting Blood Sugar	
< 120 mg/dl	258 (85.1%)
> 120 mg/dl	45 (14.9%)
resting electrocardiographic results	
normal	151 (49.8%)
having ST-T wave abnormality	4 (1.3%)
probable or definite left ventricular hypertrophy	148 (48.8%)
exercise induced angina	
No	204 (67.3%)
Yes	99 (32.7%)
slope of the peak exercise	
upsloping	142 (46.9%)
flat	140 (46.2%)
downsloping	21 (6.9%)
number of major vessels	
Major vessels:0	176 (58.1%)
Major vessels:1	65 (21.5%)
Major vessels:2	38 (12.5%)
Major vessels:3	20 (6.6%)
Missing	4 (1.3%)
results of nuclear stress test	
normal	166 (54.8%)
fixed defect	18 (5.9%)
reversable defect	117 (38.6%)
Missing	2 (0.7%)

Reference

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). rmdformats: Springer.

Barnier Documents. J (2022).rmd formats:HTML Output Formats and Templates for ‘*rmarkdown’Documents*.R package version 2.2.6, https://CRAN.R-project.org/package=rmdformats.

Malik, M., & Camm, A. J. (1990). Heart rate variability. *Clinical cardiology*, *13*(8), 570-576.

Tsuji, H., Venditti Jr, F. J., Manders, E. S., Evans, J. C., Larson, M. G., Feldman, C. L., & Levy, D. (1996). Determinants of heart rate variability. *Journal of the American College of Cardiology*, *28*(6), 1539-1546.

Kikuya, M., Hozawa, A., Ohokubo, T., Tsuji, I., Michimata, M., Matsubara, M., ... & Imai, Y. (2000). Prognostic significance of blood pressure and heart rate variabilities: the Ohasama study. *Hypertension*, *36*(5), 901-906.

Stauss, H. M. (2003). Heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *285*(5), R927-R931.

D'Antono, B., Dupuis, G., Fortin, C., Arsenault, A., & Burelle, D. (2006). Angina symptoms in men and women with stable coronary artery disease and evidence of exercise-induced myocardial perfusion defects. *American heart journal*, *151*(4), 813-819.

Kannel, W. B. (1995). Range of serum cholesterol values in the population developing coronary artery disease. *The American journal of cardiology*, *76*(9), 69C-77C.

Tan, Y. Y., Gast, G. C. M., & van der Schouw, Y. T. (2010). Gender differences in risk factors for coronary heart disease. *Maturitas*, *65*(2), 149-160.

Detrano, R. (2020, March 24). *Cleveland Clinic Heart Disease Dataset*. Kaggle. https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset/data

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.